MULTI-MODAL SENSORY FUSION WITH APPLICATION TO AUDIO-VISUAL SPEECH RECOGNITION

Stephen M. Chu and Thomas S. Huang

Beckman Institute and Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

ABSTRACT

In this work we consider the bimodal fusion problem in audio-visual speech recognition. A novel sensory fusion architecture based on the coupled hidden Markov models (CHMMs) is presented. CHMMs are directed graphical models of stochastic processes and are a special type of dynamic Bayesian networks. The proposed fusion architecture allows us to address the statistical modeling and the fusion of audio-visual speech in a unified framework. Furthermore, the architecture is capable of capturing the asynchronous and temporal inter-modal dependencies between the two information channels. We describe a model transformation strategy to facilitate inference and learning in CHMMs. Results from audio-visual speech recognition experiments confirmed the superior capability of the proposed fusion architecture.

1. INTRODUCTION

Incorporating visual information into automatic speech recognition (ASR) has been demonstrated as an effective approach to improve the performance and robustness over the audio-only systems, and has received much attention in recent years [7]. One of the most challenging issues in bimodal ASR is how to fuse the audio (i.e. acoustic speech signal) and the visual (i.e. lip motion) modalities.

The fusion of audio and visual speech is an instance of the general sensory fusion problem. The sensory fusion problem arises in the situation when multiple channels carry complementary information about different components of a system. In the case of audio-visual speech, the two modalities manifest two aspects of the same underlying speech production process. From an observer's view, the audio channel and the visual channel represent two interacting stochastic processes. We seek a framework that can model the two individual processes as well as their dynamic interactions.

One interesting aspect of audio-visual speech is the inherent asynchrony between the audio and visual channels. Most *early integration* approaches to the fusion problem assume tight synchrony between the two. However, studies have shown that human perception of bimodal speech does not require rigid synchronization of the two modalities [6]. Furthermore, humans appear to use the audio-visual asynchronies as multimodal features. For example, it is well known that the voice onset time (VOT) is an important cue to the voicing feature in stop consonants. This information can be conveyed bimodally by the interval between seeing the stop release and hearing the vocal cord vibration. Therefore, a successful fusion scheme should not only be tolerant to asynchrony between the audio and visual cues, but also be apt to capture and exploit this bimodal feature.

2. SENSORY FUSION USING CHMMS

It's a fundamental problem to model stochastic processes that have structure in time. A number of frameworks have been proposed to formulate problems of this kind. Among them is the hidden Markov model (HMM), which has found great success in the field of ASR. In recent years, a more general framework, the Dynamic Bayesian Networks (DBNs), has emerged as a powerful and flexible tool to model complex stochastic processes [3].

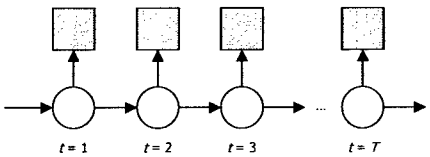

Figure 1. DBN representation of an HMM

The DBNs generalize the hidden Markov models by representing the hidden states as state variables, and allow the states to have complex interdependencies. Under the DBNs framework, the conventional HMM is just a special case with only one state variable in a time slice. DBNs are commonly depicted graphically in the form of probabilistic inference graphs. An HMM can be represented in this form by rolling out the state machine in time, as shown in Figure 1. Under this representation, each vertical slice represents a time step. The circular node in each slice is the multinomial state variable, and the square node in each slice represents the observation variable. The directed links signify conditional dependence between nodes.

It is possible to just use HMM to carry out the modeling and fusion of multiple information sources. This can be accomplished by attaching multiple observation variables to the state variable, and each observation variable corresponds to one of the information sources. Figure 2 illustrates the fusion of audio and visual information using this scheme. Because both channels share the single state variable, this approach in effect assumes the two information sources always evolves in lockstep. There-

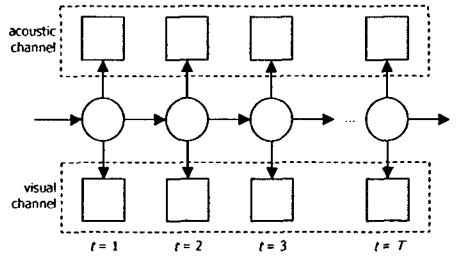

Figure 2. Audio-visual fusion using HMM

fore, it is not able to model asynchronies between the two channels.

An interesting instance of the DBNs is the so-called Coupled hidden Markov models (CHMMs). The name CHMMs comes from the fact that these networks can be viewed as parallel rolled-out HMM chains coupled through cross-time and cross-chain conditional probabilities. In the perspective of DBNs, an n-chain CHMM has n hidden nodes in a time slice, each connected to itself and its nearest neighbors in the next time slice. For the purpose of audio-visual speech modeling, we considered the case of n=2, or the 2-chain CHMMs. Figure 3 shows the inference graph of a 2-chain CHMM.

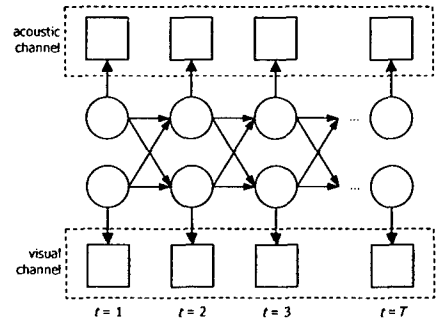

Figure 3. Audio-visual fusion using CHMM

There are two state variables in the graph. The state of the system at certain time slice is jointly determined by the states of these two multinomial variables. More importantly, the state of each state variable is dependent on both of its two parents in the previous time slice. This configuration essentially permits unsynchronized progression of the two chains, while encouraging the two sub-processes to assert temporal influence on each other's states. Note that the Markov property is not jettisoned by introducing the additional state variable and the directed links. Given the current state of the system, the future is conditionally independent of the past. Furthermore, given its two parents, a state variable is also conditionally independent of the other state variable.

In addition to the two state variables, there are two observation variables in each time slice. Each observation variable is a private child of one of the state variables. The observation variables can be either discrete or continuous. It is possible with this framework that one of the state variable is continuous and the other one is discrete.

In the context of audio-visual speech fusion, the audio and visual channels are associated with the two state variables respectively through the observable nodes. Inter-channel asynchrony is allowed. The overall dynamics of the audio-visual speech is determined by both modalities.

In general, the time complexity of exact inference in DBNs is exponential in the number of state variables per time slice. For systems with large number of state variables, exact inference quickly becomes computationally intractable. Consequently, much attention in the literature has been paid to approximation methods that aim to solve the general problem. Existing approaches include the *variational methods* [4] and the *sampling methods* [5]. However, these methods usually exhibit nice computational properties in an asymptotic sense. When the number of states is very small, the computational overhead embedded in the approximation method is often large enough to offset the theoretical reduction in time complexity. In this situation, the approximation becomes superfluous and exact inference becomes more desirable. In the following section, we describe a model transformation strategy that facilitates inference and learning in CHMMs.

3. CHMM TRANSFORMATION

The state of a 2-chain CHMM is jointly determined by the two state variables in the parallel chains. If the two state variables can take Q_1 and Q_2 discrete values respectively, then the CHMM in effect has $Q_1 \times Q_2$ possible states. The same state space can also be represented by a conventional HMM that has $Q_1 \times Q_2$ hidden states. Moreover, in CHMM, the output distribution of a joint state can be obtained by taking the product of the two output densities of the two individual state variables; Similarly, in a 2-stream HMM, the output distribution of a state is the product of the two stream-dependent densities. Hence, it is also possible to represent the output configurations of a 2-chain CHMM with a 2-stream HMM that has an equivalent state space. However, the observable nodes of a $Q_1 \times Q_2$ CHMM are fully specified by a table containing $Q_1 + Q_2$ entries. On the other hand, an unconstrained 2-stream HMM with $Q_1 \times Q_2$ hidden states has $2 \times Q_1 \times Q_2$ distinct output densities. This difference arises because in the CHMM an output node is only dependent on its single parent, while in the state-equivalent HMM the output is effectively conditioned on both state variables in the original CHMM. Fortunately, this discrepancy can be readily resolved through tying the appropriate output densities in the 2-stream HMM according to the mapping from CHMM states to HMM states. This state mapping and parameter tying procedure is easy to visualize graphically.

Figure 4 illustrates the state-machine diagram of 2-stream HMM obtained by transforming a 2-chain CHMM with $Q_1 = 3$ and $Q_2 = 2$. The state space of the original CHMM is repre-

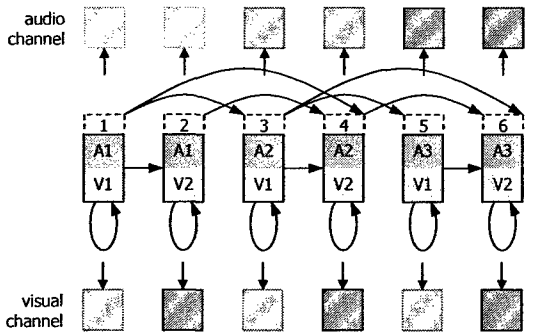

Figure 4. Transform CHMM to HMM through state-space mapping and parameter tying

sented by the 6 hidden states in the HMM. This mapping is explicitly depicted in the diagram. E.g., the state 3 in the HMM is equivalent to the state $\{q_1 = 2, q_2 = 1\}$ in the CHMM. The output densities of the HMM are tied according to the mapping. In the figure above, the observation nodes with the same color shade are tied. For example, the output densities modeling the lower stream in state 2, 4, and 6 are tied, because they all correspond to the entry $p(o_t \mid q_2 = 2)$ in the CPT of the CHMM.

The allowed state transition in the HMM is also derived from the state space mapping. In this example, it is assumed that the conditional probabilities concerning the two state variables in the CHMM satisfy the following condition.

$$P(q_{t+1}^i \mid q_t^1, q_t^2) = 0 \text{ if } q_{t+1}^i \neq q_t^i \text{ and } q_{t+1}^i \neq q_t^i + 1 \qquad (1)$$

This condition essentially enforces the left-to-right no-skip policy in the sense of conventional HMM for the two state variables in the CHMM, which is commonly used in audio-only speech recognizers. For example, a possible state path in the CHMM could be $\{q_1 = 1, q_2 = 1\} \rightarrow \{q_1 = 2, q_2 = 1\} \rightarrow \{q_1 = 3, q_2 = 2\}$, this is equivalent to the allowed state path $1 \rightarrow 3 \rightarrow 6$ in the HMM.

Other meaningful model configurations can be obtained through manipulating the allowed state transitions. For instance, it might be reasonable to model the dynamics of the lip motion using an ergodic state variable, i.e., no restriction on the possible state transitions for this variable.

It is worthy noting that the 2-stream HMM approach to audio-visual fusion as shown in Figure 2 can be considered as a special case of the CHMM-based fusion architecture. In that case, the number of the audio states must be equal to the number visual states, and the two state variables always progress in lock step, i.e. $Q_1 = Q_2$, and $q_t^1 = q_t^2$ for all t. The CHMM-based fusion architecture permits a much richer space for modeling interactions between the two modalities.

The model transformation strategy described is fairly general and can be implemented on any HMM-based ASR platforms that support multiple observation streams and parameter tying.

4. AUDIO-VISUAL ASR EXPERIMENTS

The experiments carry two objectives. The first is to evaluate the improvement in noise robustness brought by the bimodal approach to ASR. The second is to compare the performance of the proposed fusion architecture with other fusion techniques.

To fulfill the first objective, we built an acoustic speech recognizer as the baseline system. The recognizer was trained using clean speech. Noisy condition of a particular SNR level was simulated by adding white Gaussian noise to the clean speech samples. The same acoustic feature sets were also used in the audio channel of the bimodal system. However, it is assumed that visual channel is not affected by any additional noise during testing. A visual-only recognizer was built and used as a benchmark. To achieve the second objective, we implemented a common form of the early integration approach, i.e. fusion by concatenating the audio and visual feature vectors. The systems were developed using HTK.

Evaluation of the bimodal speech recognition system was performed on an audio-visual speech dataset [1] collected by Chen *et al.* at the Carnegie Mellon University. The vocabulary consists of 78 words commonly used in scheduling applications. The visual features were derived from the lip-tracking data provided with the bimodal speech dataset. The primary visual features considered in the experiments are composed of h_1, h_2, which measure the vertical openings of the upper and lower lips, and the distance between the two mouth-corners, w. Delta features were also included, thus the actual visual feature vector is six-dimensional. The acoustic speech was processed using a 25ms Hamming window, with the frame period set at 10ms. For each frame, 12 MFCC coefficients were calculated from the result of filterbank analysis using 26 channels. Delta coefficients were also computed and then appended to the static features resulting in a 24-dumentional acoustic feature vector.

We constructed the acoustic and the audio-visual speech models at the word level. The audio-only system is based on HMMs with nine states, left-to-right topology, and no skips. The HMMs used in the visual-only system have a similar topology, but with only five states. HMM configuration identical to the audio-only system is used in the early integration bimodal system. The CHMM-based bimodal system uses five states to model the audio channel and three states for the visual channel. The allowed state transitions follow the policy specified in equation (1). Recognition was performed in the connected-word mode without the help of any grammatical constrains. A cross-validation scheme was used in the evaluations due to the limited amount of data. Specifically, the recognizers were trained on a subset containing 90% of the available data and tested on the remaining 10%; this process was repeated until all data had been covered in testing. The results are summarized in Table 1.

In the recognition results, it is evident that both of the bimodal systems demonstrate improved noise robustness in comparison to the audio-only system. However, at 10dB, the gain in robustness achieved by the early integration system is very lim-

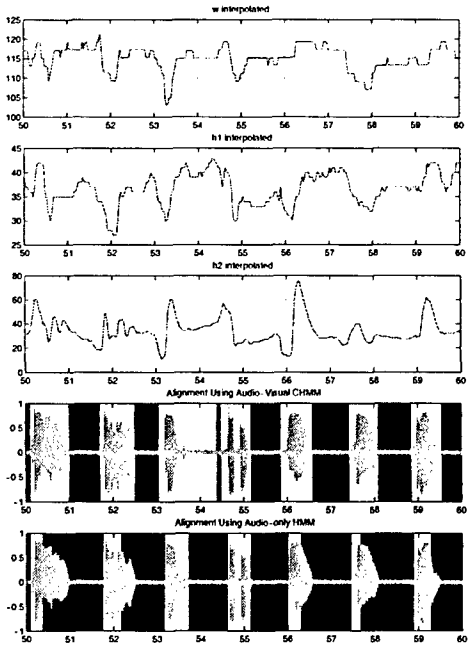

Figure 5. Forced alignment using audio only HMM and audio-visual CHMM

ited. On the other hand, the CHMM approach managed to give a clear improvement in performance at the same SNR level. At the 30dB, which is the SNR of the clean speech data, the recognition accuracy of the CHMM-based system is slightly worse than both the audio-only recognizer and the early integration bimodal system.

Table 1. Summary of recognition results (measured in %word accuracy). 'A' indicates the audio-only system; 'V' indicates the visual-only system; 'A+V' indicates the bimodal system using early integration; and 'CHMM' indicates the CHMM-based system.

SNR	10dB	20dB	30dB
A	4.03	43.61	99.10
V	42.95	42.95	42.95
A+V	10.58	72.79	99.74
CHMM	35.32	86.58	93.32

An important cue the visual modality provides in bimodal speech perception is the information about boundary locations of the speech units within an utterance. It would be interesting to see if this effect can be observed in our audio-visual ASR system. We computed forced alignment of a speech segment in the 20 dB test set using both the acoustic only recognizer and the CHMM-based bimodal recognizer. The results are illustrated in Figure 5.

Figure 5 covers a 10-second segment of the alignment result. The two subplots on the bottom show the word boundaries superimposed with the speech waveform. The upper one is the alignment obtained using audio-visual CHMMs; the lower one shows the alignment obtained using acoustic only HMMs. The three subplots on the top display the static visual features used in the bimodal system. All five plots are time-aligned so that the correspondence among them can be visualized.

From the plot, we see that the audio-only recognizer almost always give the incorrect end-of-word boundary at this noise level. In contrast, the bimodal system was able to precisely determine the end boundaries in 6 out of 7 cases. It is interesting to observe that the bimodal recognizer consistently introduced a lead-time before the audible starting point of a word. This observation is consistent with the finding from human speech perception, that the visual speech usually leads the visual speech by a varying time window. The duration of the visual lead-in shown in Figure 5 ranges from about 40ms to 150ms.

5. CONCLUSIONS

We have described a novel sensory fusion architecture based on the CHMMs. A model transformation strategy that maps the state space of a CHMM onto the state space of a classic HMM is proposed to carry out inference and learning. Bimodal speech recognition experiments demonstrate that the CHMM-based fusion scheme can utilize the information in the visual channel effectively in noisy conditions.

6. REFERENCES

[1] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18(1), pp. 9-21, 2001.

[2] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2(3), pp. 141-150, 2000.

[3] Z. Grahramani, "Learning dynamic Bayesian networks," in *Adaptive processing of temporal information* (C. L. Giles and M. Gori, eds.), Lecture notes in artificial intelligence, Springer-Verlag, 1997.

[4] M. Jordan, Z, Grahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. I. Jordan, eds. Boston: Kluwer Academic Publishers, 1998.

[5] D. J. C. Mackay, "Introduction to Monte Carlo methods," in *Learning in Graphical Models*, M. I. Jordan, eds. Boston: Kluwer Academic Publishers, 1998.

[6] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Cambridge: The MIT Press, 1998.

[7] C. Neti, *et al.*, *Audio-Visual Speech Recognition*, Final Workshop 2000 Report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, 2000.

[8] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. IEEE ICASSP*, vol. 6, pp. 3733-3736, 1998.